DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

Altruism, Reputation, and Noise in Linear Public Goods Experiments

Thomas R. Palfrey
California Institute of Technology

Jeffrey E. Prisbrey
Universitat Pompeu Fabra, Barcelona, SPAIN

SOCIAL SCIENCE WORKING PAPER 864

September 1993

Altruism, Reputation, and Noise in Linear Public Goods Experiments

Thomas R. Palfrey Jeffrey E. Prisbrey

Abstract

We report the results of a public goods experiment using a design that enables us to directly measure individual response functions in voluntary contributions games and estimate error rates. In addition, following Andreoni (1988), we employ two treatments in order to measure the extent to which voluntary contribution is due to reputation effects. The *partners* treatment involves a fixed group of subjects playing a repeated game. The *strangers* treatment approximates a one-shot game by randomly changing group assignments after each play. Our data shows that essentially the only difference between the two treatments is the amount of noise in the data, with the strangers treatment being the noisier of the two. This noise manifests itself in two distinct ways. First, there is more variation of decision rules across subjects in the strangers treatment. Second, individual behavior is, on average, less consistent with a cutpoint decision rule in the strangers treatment, which produces higher estimates of individual error rates. The differences between the strangers and partners data are virtually the same as differences between data from experienced and inexperienced subjects. This suggests an explanation for the finding in Andreoni (1988) that there was greater contribution under the strangers treatment in the standard homogeneous environment in which one direction of errors (undercontribution) are censored. Our results also support his conclusion that reputation effects do not appear to play a prominent role in repeated linear public goods voluntary contribution games. Many findings from past public goods experiments are consistent with our model of random variation in a population of subjects who are, on average, neither altruistic nor spiteful.

Keywords: Voluntary contributions, public goods, experiments, reputation, learning, errors.

JEL Classification numbers: 026, 215

Altruism, Reputation, and Noise in Linear Public Goods Experiments*

Thomas R. Palfrey Jeffrey E. Prisbrey

1 Introduction

The most common public goods experiment examines the extent to which contributions occur when individuals have a dominant strategy not to contribute. This mechanism of public good provision is called the voluntary contribution mechanism. In these experiments, a subject, who is a member of a small group, is endowed with an amount of a good that may either be consumed privately or contributed to the public good of the group. Incentives are usually designed so that a self-interested subject has a strict dominant strategy to contribute nothing, but the efficient outcome for the group is for each subject to contribute all their input to the public good.

A common finding in these experiments is that subjects often contribute, violating their dominant strategy. In addition, contribution rates have been found to be correlated with a number of treatment variables such as experience and induced preferences for the public good. However, to date there is no coherent theory that can account for the variety of findings that have been reported. A number of casual explanations for some of these findings have been offered in the literature, some suggesting a type of altruism contaminating the experimentally induced incentives,[1] and/or that the subjects are trying to establish a reputation in order to influence play later in the experiment.

In Palfrey and Prisbrey (1992) we proposed an alternative explanation, namely that most of the observed anomalies could be accounted for simply as background noise, and that the apperance of altruistic behaivor or stratigic reputation-building is illusionary or,

*We acknowledge the financial support of the National Science Foundation (SBR-9223701) and the Ministerio de Education y Ciencia (DGICYT PB91-0810). We thank Estela Hopenhayn for assistance in preparing and conducting the experiments. Antonio Rangel helped with the translation of instructions from English. We are grateful to our colleagues at both Caltech and Pompeu Fabra for their advice, with special thanks to Antoni Bosch.

[1]See the survey by Ledyard (1993)

at best, of minor importance in explaning the data. As a result of the usual experimental designs in which errors can only be manifested as overcontribution, the importance of systematic findings such as altruism and strategic play have been overstated.[2] To a limited extent, recent experiments have been conducted that lend some credence to this view,[3] but a careful study that is designed to precisely measure the relative contribtuion of each of the various proposed explanations has not yet been carried out. Unfortunately, the typical experimental designs do not permit precise measurment of the separate contribution of these diverse effects: altruism, reputation building, and noise. In this paper, we present the results of an experiment that was specifically designed to sort out these effects and accurately measure the separate contribution of each.

A basic premise of our study is that individual behavior can be decomposed statistically into a systematic component and a residual component. We call the systematic component a *decision rule*, and the residual component *noise*, or error. In the context of a linear voluntary contribution game it is natural to limit attention to very simple decision rules, called cutoff decision rules, in which an individual contributes if and only if his marginal rate of substitution between the private good and the public good is less than or equal to some critical value. This includes as a special case perfectly self-interested behavior,[4] where the critical value is 1. However, altruistic behavior or reputation-building behavior would be consistent with decision rules where the critical value is set higher than 1. "Spiteful" behavior (Saijo and Nakamura 1993) corresponds to a critical value less than 1. The noise component of individual behavior is modelled as statistical deviation from a cutpoint rule. One way to think of this is that the *observed* decision rule of a subject has some random variation over time due to extraneous factors that are essentially impossible to measure. These factors would include computational errors, errors associated with learning-by-doing, and so forth. With this interpretation of the noise component, we expect experience to lead to a decrease in noise.[5] We interpret such decreases in noise as evidence of *learning*.

Past experimental designs make it virtually impossible to accurately identify the decision rule component from the noise component. In those experiments, there is little if any variation of the marginal rate of substitution. Typically, everyone has the same marginal rate of substitution throughout the experiment, and it is greater than 1. The focus of attention is on the aggregate frequency of violations of a deterministic version of the self-interest model of behavior. In the context of our non-deterministic two-component model of individual behavior, contribution could be due to altruism or reputation building or it could be due to noise. In those experiments, noise leads to systematic bias in the data, in that (at least relative to the self-interested model) only noise that leads to

[2]Overcontribution is small in magnitude or nonexistent in other public goods experiments where errors can be made in both directions. See Palfrey and Rosenthal (1988, 1991), and the references they cite.

[3]See, for example, Andreoni (1988, 1992) and Saijo and Nakamura (1993).

[4]Reputational play could also involve more complicated decision rules where the cutpoint changes over time or as a function of history.

[5]Experience could also lead to adaptation of the decision rule, although we find little evidence for this.

contribution can possibly be observed.

An accurate measurement of a subject's decision rule and the magnitude of the noise component is possible in a heterogeneous and changing environment; an environment where a subject faces a number of different marginal rates of substitution, and yet his information is otherwise the same. It is then possible, by a variety of methods (Palfrey and Rosenthal 1991), to estimate the subject's decision rule. As well as estimating the extent to which cutpoint rules deviate from 1, these methods also calibrate the noise component. The design reported here systematically varies each subject's marginal rate of substitution in order to estimate the distribution of decision rules and the distribution of the error rates. This allows us to measure the extent to which altruism or reputation builiding explains the commonly observed overcontribution and the extent to which these observations can be accounted for simply as noise. This also allows us to measure the extent to which players learn with experience.

Once the noise component and the systematic component of individual choice behavior have been separated, the next step is to break down the systematic component of decision rules and identify the relative importance of altruistic behavior and strategic reputation-building behavior. Following the approach of Andreoni (1988), we do this by conducting half of the experimental contribution games as a sequence of one-shot encounters with changing group memberships (the "strangers" treatment) and half the contribution games as a sequence of encounters where group membership remains fixed (the "partners" treatment).

The difference between the decision rule in a series of one time encounters and the decision rule in a similar number of encounters repeated within the same group could be attributed exclusively to reputation building. Accordingly a comparison between the decision rules measured under the two treatments is then made. If reputation-building is an important part of the explanation, we should observe decision rules with higher critical points in the partners treatment than in the strangers treatment. In addition, we should observe significantly more decay (declining contribution rates over the course of an experiment) in the partners treatment. The ability of our method to measure error rates means that we are able to draw firm conclusions about whether decay in previous experiments was due to learning or was evidence of reputation building.

The rest of the paper is organized as follows. Section 2 discusses the relevent findings from past experiments. Section 3 describes our experimental environment and the details of the design. Section 4 analyzes the data. We make concluding remarks in section 5.

2 Previous Research

The experimental study of public good provision by the voluntary contribution mechanism has a history that is well detailed in Dawes (1980) and in Ledyard (1992). Almost all past research, including the influential works of Marwell and Ames (1979, 1980, 1981),

Issac and Walker (1988, 1989), Issac, Walker and Thomas (1984), and Andreoni (1988), examine situations in which each subject's marginal rate of substitution is fixed for all periods of the experiment, usually all subjects are assigned identical valuations.

A number of general findings have emerged from the literature:

- aggregate contribution rates range between 20 percent and 50 percent,
- at some point in time and in violation of dominant strategy incentives, nearly all players contribute to the public good,
- there is a strong negative relationship between the marginal rate of substitution and the rate of contribution, and
- contribution rates fall with repetition and with experience (where repetition represents a sequence of decisions within the same group, and experience represents another similar sequence of decisions with a different group).

And, with regard to learning and reputation effects, Andreoni (1988) finds that:

- subjects in repeated encounters contribute less to the public good than subjects in one time encounters,
- the proportion of *free riders*, or subjects that consistently use the dominant strategy decision rule, is greater in repeated encounters than in one time encounters, and,
- experience effects are greater for subjects in one-time encounters than for subjects in repeated encounters.

A number of papers (Ledyard, 1993 and its references) have tried to attribute the contributions to altruism on the part of the subjects. It is argued that the experimentally induced monetary incentives do not fully control for all aspects of a subject's utility, and that utility may partly depend on the welfare or efficiency of the group outcome as well as monetary payoff. If the amount of consideration given to the group outcome is high enough, contribution to the public good is consistent with utility maximization.

On the other hand, the presence of altruism does little to explain the counter-intuitive results in Andreoni (1988). After all, with the additional assumption of incomplete information, the ability to establish reputations is known, at least theoretically, to justify the use of dominanted strategies, see Kreps, et al. (1982). The work of Kreps, et al. suggests that, if anything, the contribution rates in repeated encounters should be higher, not lower, than the contribution rates in one time encounters.

In addition to the systematic qualitative features of the data noted above, there is also a lot of statistical variation across trials. This suggests yet another explanation

which is simply that the data is noisy.[6] Because of the experimental designs that are used, "noise" (in the sense of statistical deviation from the theoretical prediction) can only manifest itself as contribution. None of the past studies are designed to collect data that enables accurate measurement of the seperate effects of "noise" and "altruism" on voluntary contributions.

Recently, Andreoni (1992) and Palfrey and Prisbrey (1992) have designed experiments that do enable the differentiation. Andreoni proceeds by comparing data collected from a standard environment with data collected from a similar, environment in which group efficiency shall no longer be important to the individual. Andreoni attributes actions which help the group in the manipulated environment to "confusion," and he attributes the additional contribution in the standard environment to altruism.

Building on Palfrey and Prisbrey (1992) we use a heterogeneous environment, in which each individual's marginal rate of substitution is varied over the course of an experiment. By observing a subject's decisions at a number of different marginal rates of substitution, instead of at just one, and by assuming that subjects make errors at some non-negative rate (possibly zero), the subject's entire response function can be estimated. Using the separate techniques of probit and classification analysis, they are able to directly measure the rate of errors in the subject pool, and also directly measure contributions due to altruism.

The research presented reexamines the surprising partners-strangers findings of Andreoni (1988) in the heterogeneous environment of Palfrey and Prisbrey (1992), and proposes an explanation consistent with his findings and findings in past experiments. This new explanation combines the "uncontrolled incentives" rationalization with a statistical model of subject decision errors. The design permits a separation of the three basic effects that have been hypothesized to explain voluntary contribution in experiments; altruism, reputation building, and noise. It also allows direct measurement of experience effects.

3 The Independent Private Values Environment

Consider a group of N individuals, each with X_i, a divisible endowment of a private good, and a value for increments of the private good. Each individual must choose an amount of their endowment to keep and an amount to give to the public good. The utility of the individual is

$$U(y, x_i) = Vy + r_i x_i,$$

where V is the value of the public good, y is the amount of the public good produced by the entire group, r_i is the individual's value for the private good, and x_i is the amount of

[6] One can imagine many reasons for the data to be noisy: incomplete subject understanding of the rules; low payoff salience; boredom; experimenter effects; demand effects; etc.

the endowment that is kept for private use. The technology is such that, for every unit of the private good contributed, one unit of the public good is produced.

By varying an individual's r_i over a number of decision periods, it is possible to estimate that individual's decision rule, $D_i(r_i/V)$, where r_i/V is the individual's marginal rate of substitution. Theoretically, an individual's decision rule should be of the following form:

$$D_i(r_i/V) = \begin{cases} 0 & \text{if } r_i/V < 1 + a_i + \varepsilon_i \\ X_i & \text{otherwise.} \end{cases},$$

where a_i is individual i's level of altruism, and ε_i is a random error term. This type of decision rule is called a *cutpoint rule* and the value $c_i = 1 + a_i$ is called the *cutpoint.* Without the error term ε_i and as long as the game does not have an infinite number of decision periods, the above rule is the complete information dominant strategy decision rule. The inclusion of the error term accounts for the possibility of random errors or unpredictable behavior by subjects.

Depending on the assumptions made about the distributions of a_i and ε_i, it is possible to estimate the decision rules in a variety of ways. Possible assumptions about a_i are: all individuals are have the same level of altruism and therefore the same a_i, a_i is never negative, or a_i is drawn from some distribution. There are also many ways in which ε_i can be distributed, some which assume all types of errors are equally likely and others which assume that drastic errors are less likely.

We will offer two methods for estimating the decision functions. The first is to use an ordered probit analysis. The ordered probit analysis implicitly assumes that all the subjects use the same decision rule and that the ε_i's are distributed in a Normal distribution with mean zero. The assumption of a Normal distribution makes drastic errors, (contributing when r_i/V is much larger than c_i), less likely than small errors (contributing when r_i/V is close to c_i). The second method is non-parametric and is called a classification errors analysis. This method is used for the estimation of individual decision rules.

4 Experimental Design

All experiments were run using computers in the experimental economics lab at the Universitat Pompeu Fabra. A copy of the instructions is in the appendix. The data was generated by four experimental sessions each containing four experiments. There were twelve first year, undergraduate, economics students who participated in each session, making a total of forty-eight.

In all experiments, individuals were assigned to groups of four. They were given an endowment in the form of nine tokens and were told that they must choose to either

spend or keep these tokens. The subjects were then told how much each token was worth to them, in points, if they kept it. This value, which we called the *token value*, was randomly drawn from a Uniform distribution over the interval $[1, 20]$. Each token value in the group was the result of a different draw from the distribution, so group members likely had different token values—a fact which was carefully explained in the instructions. The token value is equivalent to r_i. The subjects were also told that every member of their group, including themselves, would earn a specified amount (V, in our notation) for every token they spent.

The subjects were told that they would be in four experiments, each of which had ten decision periods. The specified amount V depended upon the particular experiment; in the first two experiments of each session it was six points, in the last two experiments of each session it was ten points.

In two of the sessions, which, following Andreoni (1988) we call *Strangers*, the subjects were randomly assigned new groups after each decision period. The random assignment process was used to approximate one time encounters. In the other two sessions, named *Partners*, the subjects were assigned to new groups only between each of the four experiments. During a particular ten period experiment, the Partners were repeatedly in the same group. The subjects, of course, were told if their groups would be randomly changed between periods or if they would remain the same between periods.

The subjects were then told the rate of exchange between cash and points, and quizzed before the experiments were run. After the experiments, each subject was paid anonymously in cash.

This design enables us to examine experience effects, in addition to the effects of partnership. All the decisions in the first and the third experiment of each session are coded as inexperienced decisions. The rationale for this division is that in the first and third experiments, the subjects see a particular public good value for the first time. In the second and fourth experiments in each session, the subjects see a public good value for the second consecutive time, and these decisions are coded as experienced. No subject participated in more than one session.

5 Analysis of the Data

The data analysis centers on the measurement of subject decision rules and is specifically organized around the measurement of *cutpoint rules* and *error rates.*

5.1 Aggregate Data – A Simple Classification Analysis

As the first cut at measurement, we estimate a common cutpoint, c, and common error rate, ε, which best describes the aggregate data. The analysis proceeds by determining

the rate of classification errors in the data for each possible cutpoint. A decision is classified as an error if, under the hypothetical cutpoint rule, a subject was supposed to contribute a token, *i.e.* the subject had a value r_i/V which was strictly less than the hypothetical cutpoint, and the subject did not contribute the token. A classification error can also occur if, under the hypothetical cutpoint rule, the subject was not supposed to contribute a token and the subject did contribute the token. The estimated common cutpoint, c^*, is the one with the least classification errors and the estimated common error rate, ε^*, equals the rate of classification errors if c^* is the cutpoint.

Figures 1 and 2 show the number of classification errors as a function of the hypothetical cutpoint, and illustrate the effects of reputation and experience, respectively. In every case, the theoretical cutpoint with the lowest rate of classification errors is $c^* = 1$, which is consistent with the joint hypothesis of (a) homogeneity of subject decision rules and (b) no altruism in the subject pool. Based on this sample analysis, subjects maximize utility, and occasionally make errors.

Next consider the hypothesis suggested by the reputational model, that subjects in one time encounters have a lower cutpoint than subjects in repeated encounters. Figure 1 shows that the c^* in the Strangers condition is equal to the c^* in the Partners condition, so the reputation hypothesis is not supported. Using this method of decision rule estimation, there is absolutely no evidence of a reputation effect.

On the other hand, the data show support for an alternative "noise" hypothesis to account for the differences between the strangers and partners data: subjects in one time encounters have a higher error rate than subjects in repeated encounters. The data also show that experience reduces error rates (Figure 2).

The graphical presentation is further reinforced by a least squares regression with the average group error rate per round, assuming that all subjects use a cutpoint of 1, as the dependent variable. The regression contains four independent variables: a constant; PART, which is 1 for data from the Partners treatment and 0 for data from the Strangers treatment; EXPER, which is 1 for data from experienced subjects and 0 otherwise; and PER, which runs form 1 to 10 and is the number of the period. The results of the regression are in Table 1.

The variable PART is negative and significant, reflecting the lower average error rates in repeated encounters. The variable EXPER is also negative and significant, reflecting the lower error rates in experiments with experienced subjects. The regression also shows that error rates fall over a 10 round session since the coefficient on PER is negative and, for a one-tailed test, significant.

5.2 Aggregate Data – An Ordered Probit Analysis

An alternative approach is ordered probit analysis (McKelvey and Zavonia, 1975). Again, homogeneity of subject decision rules is assumed i.e., every subject has the same cutpoint

and error rate. The difference is that the error term, ε_{it} is assumed to be independently distributed across periods and subjects with a Normal distribution with a mean of zero and a variance of one. Hence contributing when r/V is much larger than c^* is less likely than contributing when r/V is near to c^*.

The ordered probit analysis estimates the probability of any number of tokens being contributed as a function of the marginal rate of substitution. It is easy to measure the impact of reputation, experience and time with the addition of dummy variables and a time trend. The dependent variable in the analysis, then, is the subject's decision, a number from 0 to 9. The independent variables are: a constant; r/V; PART and PARTS which are, respectively, constant and slope[7] dummies for the partners treatment; EXPER and EXPERS which are, respectively, constant and slope dummies for experience effects; and LATE and LATES which are, repsectively, constant and slope dummies for decay effects[8] over a 10-period session.

We calculate a probit response curve equal to the predicted percentage of tokens contributed as a function of r/V and plot this curve for several of the treatments in Figure 3. To do this we first compute a "score" for each value of r/V, that determines the location of the mean of a Normal density function on a line divided into intervals by the probit-generated threshold values. In the present situation there are nine intervals, one interval for each of the possible decisions, 0–9. The area under the density and between the thresholds n and $n-1$ is equivalent to the estimated probability that event n occurs. A curve which gives the expected contribution as a function of r/V can then be generated.

It should be noted that changes in the coefficients of the slope dummies will change the steepness of the expected contribution curve as well as its position relative to the x-axis, whereas changes in the coefficients of the intercept dummies will change only the position of the resulting curves with regard to the x-axis.

Recall that, theoretically, the subjects should use a cutpoint decision rule. There should be an c such that the probability of contribution if $r/V < c$ is one and the probability of contribution if $r/V > c$ is zero. If the subjects behaved in a way that was perfectly consistent with the theory and made no errors, their decision rule would be graphed as a step function which stepped from nine to zero at the cutpoint.

If the subjects do not adhere perfectly to a cutpoint decision rule the estimated curve would not be a step function, but would be S shaped. The more mistakes that were made, the flatter the curve would become. The cutpoint would be equal to the point at which they were indifferent between contributing and keeping their tokens or where the probability of contribution is 0.5. The expected contribution at the cutpoint, then, should be equal to half of the endowment, or 4.5 tokens.

[7]Slope dummies are the product of the dummy variable and r/V.

[8]Recall that past experiments have observed that contribution rates decay over a 10-period session. The dummy variable LATE is 0 in rounds 1–5 and 1 in rounds 6–10.

Refer again to Figure 3. Notice the proximity of the estimated cutpoints in each treatment. The cutpoints from the inexperienced treatments are almost identical as are the cutpoints from the experienced treatments. Furthermore, all four cutpoints are very close (within .05) to one and all four response curves intersect at one point, where $MRS = 1$. The main difference between the curves is in their slopes. The steepest curve comes from the Partners with experience treatment, the next from the Partners with no experience, the next from the Strangers with experience, and the flattest curve is from the Strangers with no experience treatment. These observations are consistent with the results of the previous section.

Both variables PART and PARTS are significant, but the significance does not support the hypothesis that repeated play leads to greater contribution. In both the experienced and the inexperienced treatment, the subjects in one time encounters have a higher cutpoint than subjects in repeated encounters, although the difference is small in magtitude, less than .05 percent.

The fact that PARTS is significant indicates that there is more noise in the one-shot treatments then in repeated encounters. The subjects in one time encounters have flatter expected contribution curves and therefore have a higher error rate. The variable EXPERS is significant and lends support to the hypothesis that experience reduces noise. The inexperienced subjects have flatter expected contribution curves and therefore have a higher error rate. The experience subjects also have a slightly lower cutpoint. The positive coefficient on EXPER compensates for this downward shift, so that the estimated cutpoint for the experienced subjects is almost identical to the inexperienced subjects' cutpoint. The coefficients on LATE and LATES mirror these results, indicating that the effect of the 10-period repetition is similar to experience effects. The response curves are steeper in the last half of a 10-period session than in the first half, but the overall contribution rate is essentially unchanged.[9] At first glance, this would seem to contradict past findings of significant delay. But in fact, there is no contradiction at all. It simply means that the observed decay in past experiments was due to learning, not reputation, a finding that was observed by past designs in which over contribution due to error is difficult to separate from over contribution for other reasons.[10]

This lack of reputation effects is further illustrated in Table 3, where the effect of PART and LATE on average contributions is cross-tabulated. Reputation effects would predict *more* decay in the partners treatment than in the strangers treatment. In fact, the opposite is observed (although the difference is not statistically significant at the 5 percent level).

Finally, consider the ordered probit model using only data from the last rounds of every experiment (Table 2). Theoretically, there should be no difference between the

[9]The average contribution rate in rounds 1–5 is 3.583 and the average contribution rate in rounds 6–10 is 3.585.

[10]If we censor all our observations with MRS$<$ 1, then indeed we also measure significant decay that is large in magnitude.

partners and strangers in these rounds. There is no chance for further reputation building. This is what we find. None of the variables outside of the constants and r/V are significant, and all except PARTS have coefficients that are much smaller in magnitude.

5.3 Individual Data – Classification Analysis

The analysis above is carried out under the maintained hypothesis of homogeneity of subject decision rules.

In this section, we apply the simple classification analysis of section 5.1 at the individual level. By doing so, we are able to estimate a *distribution*[11] of cutpoints across the entire subject pool. From these estimated cutpoints, we can compute error rates for each individual as the percentage of decisions that violate their estimated cutpoint rule. The distributions of error rates and the distribution of cutpoints are then compared across treatments.

5.3.1 The distribution of individual cutpoints

Figure 4 displays the distribution of estimated individual cutpoints across the 192 observations.[12] It is clear from this figure that the distribution[13] is centered at 0 (i.e., Nash cutpoints) and is nearly symmetric. The median cutpoint is 0 and accounts for approximately 30 percent of the observations. Two thirds of the observations range from -3 to $+3$, with the remaining one-third evenly divided below -3 and above $+3$. Three quarters of the observations range from -4 to $+4$, again with the remainder being evenly divided between large negative and large positive cutpoints.

If we break down the distribution of cutpoints by the partners/strangers treatment, we find a systematic effect, but not what one would expect from the hypothesis that repeated groups have "reputation effects" that lead to more contribution. The reputation hypothesis predicts that repeated groups will have cutpoints that are typically higher than the cutpoints in the one-shot treatment. *We do not find this.* The average or median cutpoint in both treatments equals 0. The difference between the two distributions is that the distribution for strangers is more dispersed than the distribution for partners. This is illustrated in Figure 5 which displays the empirical cumulative frequencies separately for the strangers data and the partners data. As one can see, the distribution for

[11] Rapoport (1987) has argued that heterogeneity may be an important ingredient of a complete explanation for behavior in other (step-level) public goods environments. Isacc, Walker, and Thomas (1984) and Ledyard (1993) make similar points.

[12] For each of our 48 subjects we report four separate "observations" corresponding to the four treatments that a subject participated in: low-V-inexperienced, low-V-experienced, high-V-inexperienced, and high-V-experienced. Recall that each subject participated in four different groups of four, the first two with one level of V and the second two with another level of V.

[13] Deviations from Nash cutpoints are measured in token value units. A cutpoint of 0 corresponds to $MRS = 1$ in earlier figures. In a few of the observations, more than one hypothetical cutpoint minimized classification errors. Such ties were broken by choosing the one closest to 0.

strangers looks like a mean-preserving spread of the distribution for partners. Also, both distributions are symmetric about the Nash prediction of a 0 cutpoint.

5.3.2 The distribution of classification errors

From the above classification analysis, we can obtain estimates for the distribution of classification errors across individuals. Figure 6 displays the empirical distribution of error rates across the 192 observations, where the error rate is computed as the fraction of an individual's decisions (within one treatment) that are misclassified according to that individual's estimated cutpoint. Figures 7 and 8 show the effect of experience and partnership, respectively. The effect of experience is very similar to the effect of partnership. In both cases, there is a leftward shift in the error rate distribution, indicating fewer errors with experience, and fewer errors in the partners treatment than in the strangers treatment.

There is also evidence from the joint distribution of error rates and cutpoints indicating that the distribution of classification errors may be less dispersed than Figures 4 and 5 suggest. Figure 6 displays the average error rates as a function of how far the estimated cutpoint is from 0 (perfect Nash behavior). The error rates are sharply increasing as a function of deviation from Nash play. Subjects who are estimated as Nash players have an error rate less than .05, while subjects that have a one token-value unit deviation from Nash play have twice that error rate. Observations of 5 or more unit deviations from the Nash cutpoint have more than triple that error rate. A possible explanation for this is that subjects with high estimated deviations from Nash play have high estimated error rates because they are learning, and adjusting their cutpoint decision rule over time. This is hard to verify directly since we do not have enough data to estimate reliably a trend in individual cutpoint rules. Such a finding could also be the (spurious) result of heterogeneity in individual error rates, since in small samples subjects with high error rates will produce cutpoint estimates with higher variance.

6 Conclusions

The results in this paper point to a new interpretation of observed violations of dominant strategies to free ride in voluntary contributions experiments. The explanation we suggest is not that subjects are on average either particularly altruistic nor particularly spiteful. Furthermore, consistent with Andreoni (1988) we find no evidence of reputational effects of the sort proposed in Kreps and Wilson (1982). Rather, subjects exhibit statistical fluctuations in their decision making, that manifests itself as random noise[14] in the data. This explanation is consistent with the analysis we conduct, both at the aggregate level and at the individual level.

[14]Presumably these statistical fluctuations are not purely random from the point of view of a subject making the decision.

How does such an explanation account for the apparently altruistic behavior in past experiments where subjects have a dominant strategy to free ride? The answer is simple. In those experiments, the design automatically censors all observations of subjects who have a dominant strategy to give, but end up free riding. In other words, in past experiments, the only kind of "error" relative to Nash theory that could be observed was seemingly altruistic behavior. If one re-examines our data *censoring all observations of* $MRS < 1$ (dominant strategy to give), then one finds aggregate contribution rates that are statistically significant, and of a magnitude comparable to what has been found in these other studies. Moreover, as in Andreoni (1988) we find more contribution in the strangers treatment than in the partners treatment. We are able to show that this difference is due to factors affecting the *variance* in subjects' decisions and decision rules, not a systematic tendency of *mean* behavior away from the Nash equilibrium. A similar explanation applies to the Saijo and Nakamura (1993) experiments where subjects have a dominant strategy to give in, but substantial free riding is observed. The observation that experience reduces violations is just a manifestation of experience producing lower error rates and lower subject variation.

Appendix

INSTRUCTIONS

This is an experiment in decision making. You will be paid *in cash* at the end of the experiment. The amount of money you earn will depend upon the decisions you make and on the decisions other people make. We request that you do not talk at all or otherwise attempt to communicate with the other subjects except according to the specific rules of the experiment. If you have a question, feel free to raise your hand. One of us will come over to where you are sitting and answer your question in private.

The session you are participating in is broken down into a sequence of four separate experiments. Each experiment will last 10 rounds. At the end of the last experiment, you will be paid the total amount you have accumulated during the course of all 4 experiments. Everyone will be paid in you in private and you are under no obligation to tell others how much you earned. Your earnings are given in *points*. At the end of the last experiment, you will be paid 10 pesetas for every 100 *points* you have accumulated during the course of all four

In each experiment you will be divided into 3 groups of 4 persons each. Those groups will stay the same for all 10 rounds of the experiment. After each 10 round experiment, everyone will be regrouped into 3 entirely new groups. Therefore, whenever we change groups, the other people in your group will be different from the last group you were in. You will not be told the identity of the other members in your group. Since we will be running 4 experiments tonight, you will be assigned 4 different groupings, one for each 10 round experiment.

Each round of the experiment you will have 9 tokens. You must choose how many of these tokens you wish to keep and how many tokens you wish to spend. The amount of money you earn in a round depends on how many tokens you keep, how many tokens you spend, and how many tokens are spent by others in your group. Each round, you will be told how many *points* each token is worth if you keep it. This amount is called your TOKEN VALUE and it will change from round to round and will vary from person to person randomly. To be more specific, in each round, your token value is equally likely to be anywhere from 1 to 20 *points*. There is absolutely no systematic or intentional pattern to your token values or the token values of anyone else. The determination of token values across rounds and across people is entirely random. Therefore, everyone in your group will generally have different token values. Furthermore, these token values will change from round to round in a random way. You will be informed PRIVATELY what your new token value is at the beginning of each round and you are not permitted to tell anyone what this amount is.

After being told your token value, you must wait at least 10 seconds before making your decision of how many tokens to spend and how many to keep. Your keyboard will be frozen for this period of time. When everyone has made a decision, you are told how

many tokens were spent in your group and what your earnings were for that round. This will continue for 10 rounds. Following each round you will begin with 9 new tokens and you will be randomly assigned a new token value between 1 and 20 *points*.

PAYOFFS

You will receive 3 *points* times the total number of tokens spent in your group. In addition, you will also receive your token value times the number of tokens you keep. Notice that this means every time anyone in your group spends a token, everyone in the group (including the spender) gets an additional 3 *points*, but the spender forgoes his or her token value for that token. WHAT HAPPENS IN YOUR GROUP HAS NO EFFECT ON THE PAYOFFS TO MEMBERS OF THE OTHER GROUPS AND VICE VERSA. Therefore, in each round, you have the following possible earnings, as shown in the table:

[WRITE EARNINGS TABLE ON BOARD]

Earnigs Table

YOUR SPENDING DECISION	OTHERS	YOUR EARNINGS (in *points*)
0	N tokens	(N*3) + (9*your token value)
1	N tokens	3 + (N*3) + (8*your token value)
2	N tokens	6 + (N*3) + (7*your token value)
3	N tokens	9 + (N*3) + (6*your token value)
4	N tokens	12 + (N*3) + (5*your token value)
5	N tokens	15 + (N*3) + (4*your token value)
6	N tokens	18 + (N*3) + (3*your token value)
7	N tokens	21 + (N*3) + (2*your token value)
8	N tokens	24 + (N*3) + your token value
9	N tokens	27 + (N*3)

Here is an example: Suppose everyone else in your group spends 13 tokens in all and you spend 4 tokens and your token value was 12. You would earn $12 + 39 + 60 = 111$ *points*. If you had spent 3 tokens you would have earned $9 + 39 + 72 = 120$ *points*. If you had spent 5 tokens you would have earned $15 + 39 + 48 = 102$ *points*.

Are there any questions? [ANSWER QUESTIONS]

[Two practice rounds. Tell them not to press any keys unless you tell them to. In round 1 have each subject spend the number of tokens equal to the last digit of their ID#. In round 2 have each subject KEEP the number of tokens equal to the last digit of their ID#. Go over screen display and history. Tell subjects to refrain from pressing keys for no reason.]

[Keep screen display on]

[Hand out quiz.]

[Correct quiz answers and read them aloud.]

[Answer any additional questions.]

[Begin experiment 1.]

Specific instructions for Experiment 2:

Experiment 2 is the same as experiment 1 except now everyone in a group receives 6 *points* times the number of spenders in their group. Again, in addition, nonspenders also receive their token values. Again, everyone has been reassigned to a new group with a new set of participants. Here is your new payoff table:

[CHANGE BOARD. EXPLAIN.]

Example: Suppose everyone else in your group spends 13 tokens in all and you spend 4 tokens and your token value was 12. You would earn $24 + 78 + 60 = 162$ *points*. If you had spent 3 tokens you would have earned $18 + 78 + 72 = 168$ *points*. If you had spent 5 tokens you would have earned $30 + 78 + 48 = 156$ *points*.

Specific instructions for Experiment 3:

Experiment 3 is the same as experiments 1 and 2 except now everyone in a group receives 10 *points* times the number of spenders in their group. Again, in addition, nonspenders also receive their token values. Again, everyone has been reassigned to a new group with a new set of participants.

[CHANGE BOARD. EXPLAIN.]

Example: Suppose everyone else in your group spends 13 tokens in all and you spend 4 tokens and your token value was 12. You would earn $40 + 130 + 60 = 230$ *points*. If you had spent 3 tokens you would have earned $30 + 130 + 72 = 232$ *points*. If you had spent 5 tokens you would have earned $50 + 130 + 48 = 228$ *points*.

[Begin experiment 3.]

Specific instructions for Experiment 4:

Experiment 4 is the same as experiments 1, 2 and 3 except now everyone in a group receives 15 *points* times the number of spenders in their group. Again, in addition, nonspenders also receive their token values. Again, everyone has been reassigned to a new group with a new set of participants.

[CHANGE BOARD. EXPLAIN.]

Example: Suppose everyone else in your group spends 13 tokens in all and you spend 4 tokens and your token value was 12. You would earn $60 + 195 + 60 = 315$ *points*. If you had spent 3 tokens you would have earned $45 + 195 + 72 = 312$ *points*. If you had spent 5 tokens you would have earned $75 + 195 + 48 = 318$ *points*.

[Begin experiment 4.]

Regression on Average Errors	
one	0.24140
	(15.24312)
PART?	-0.03541
	(-2.88723)
EXPER	-0.03287
	(-2.67963)
PER	-0.00403
	(-1.88879)
No. of Obs.	160
R^2	0.10900
$\bar{R}^2$	0.09186

Table 1: A least squares regression with the average error rate per round, across subjects as the dependent variable.

	All Rounds	Last Round
one	1.57	1.80
	(16.27)	(6.01)
r/V	-.66	-.81
	(-11.00)	(-4.42)
PARTS	-.18	-.18
	(-2.81)	(-0.84)
PART	.17	.03
	(1.73)	(.08)
EXPERS	-.15	-.11
	(-2.37)	(-.52)
EXPER	.15	.01
	(1.48)	(.04)
LATES	-.18	
	(-2.86)	
LATE	.18	
	(1.76)	
λ_1	.29	.31
	(19.61)	(6.38)
λ_2	.55	.59
	(38.17)	(14.16)
λ_3	.77	.76
	(59.52)	(20.93)
λ_4	.93	.90
	(75.23)	(28.61)
λ_5	1.09	.99
	(98.05)	(33.66)
λ_6	1.19	1.09
	(105.62)	(31.52)
λ_7	1.35	1.26
	(95.63)	(27.25)
λ_8	1.57	1.49
	(74.11)	(20.87)
ln lik	-3303.2	-325.17
N	1920	192

Table 2: Ordered Probit Analysis: The dependent variable is the number of tokens contributed. Under each coefficient is the asymptotic t-statistic. The log likelihood and sample size are also given.

	partners	strangers
early ($t = 1 - 5$)	3.39	3.78
late ($t = 6 - 10$)	3.53	3.64

Table 3: Mean contribution (out of nine tokens) as a function of LATE and PART. $N = 480$ in each cell.

CUTPOINT ANALYSIS
UPF Data
Partners vs. Strangers

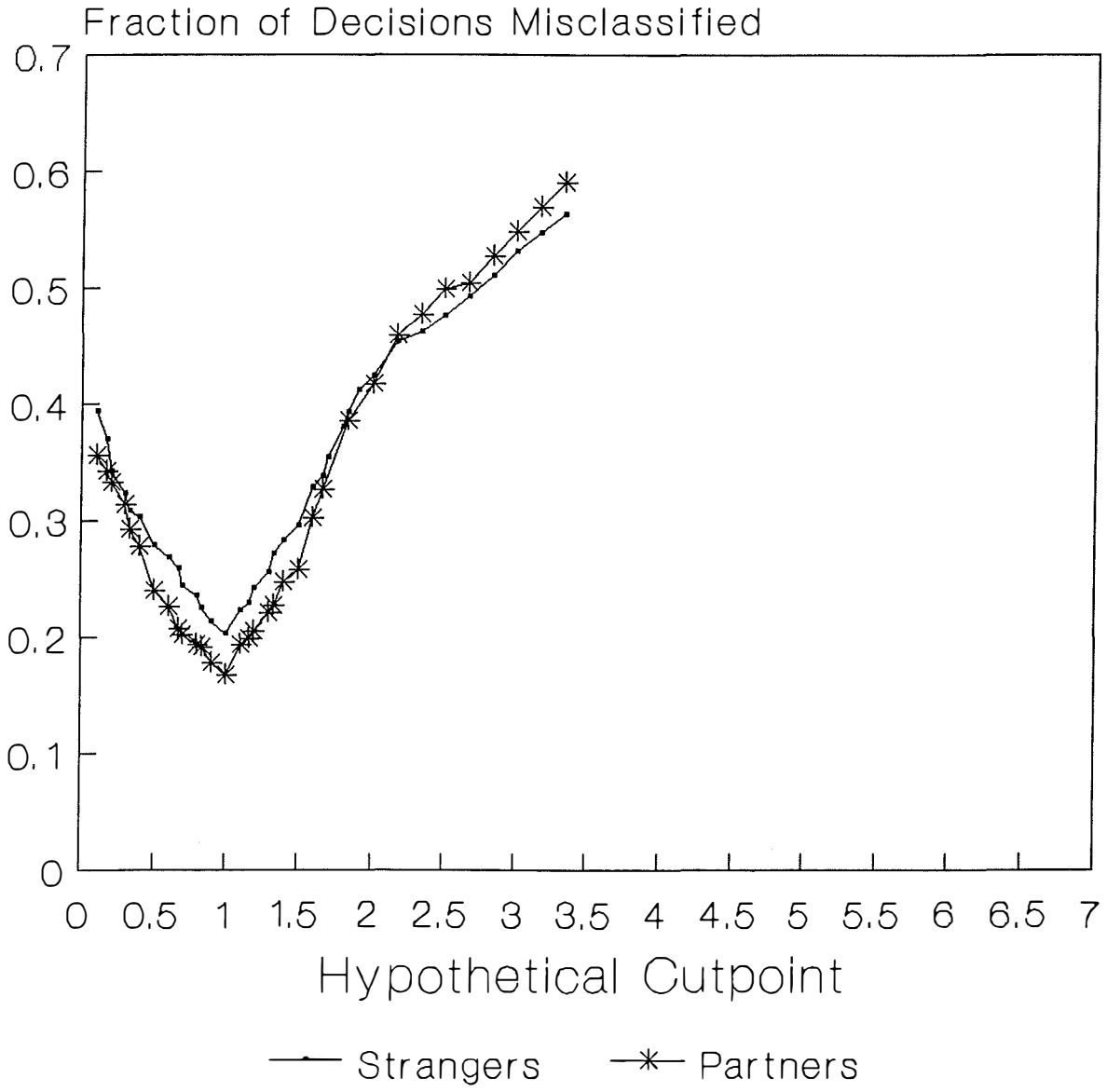


Figure 1

CUTPOINT ANALYSIS
UPF Data
Experience Effects

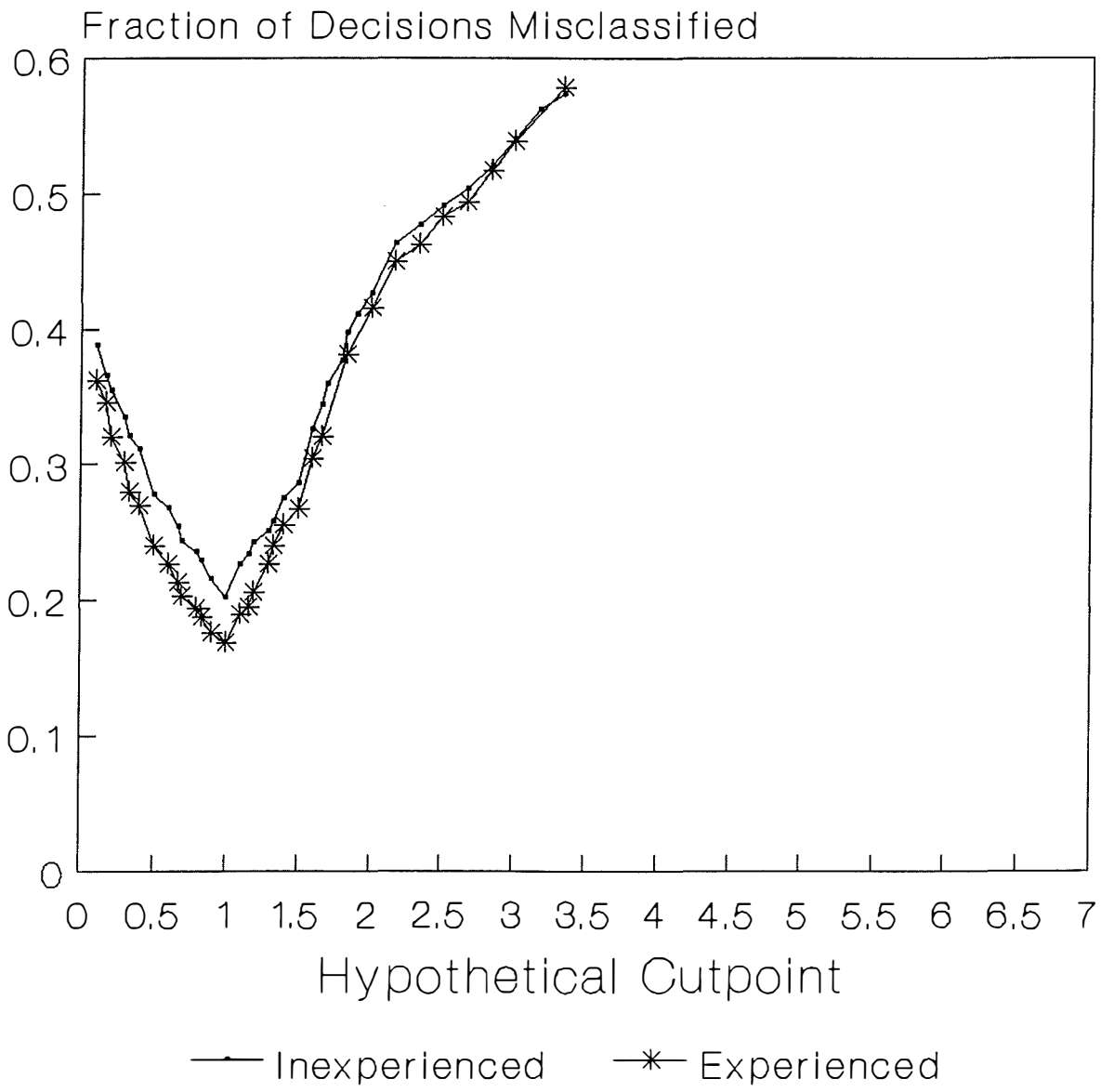


Figure 2

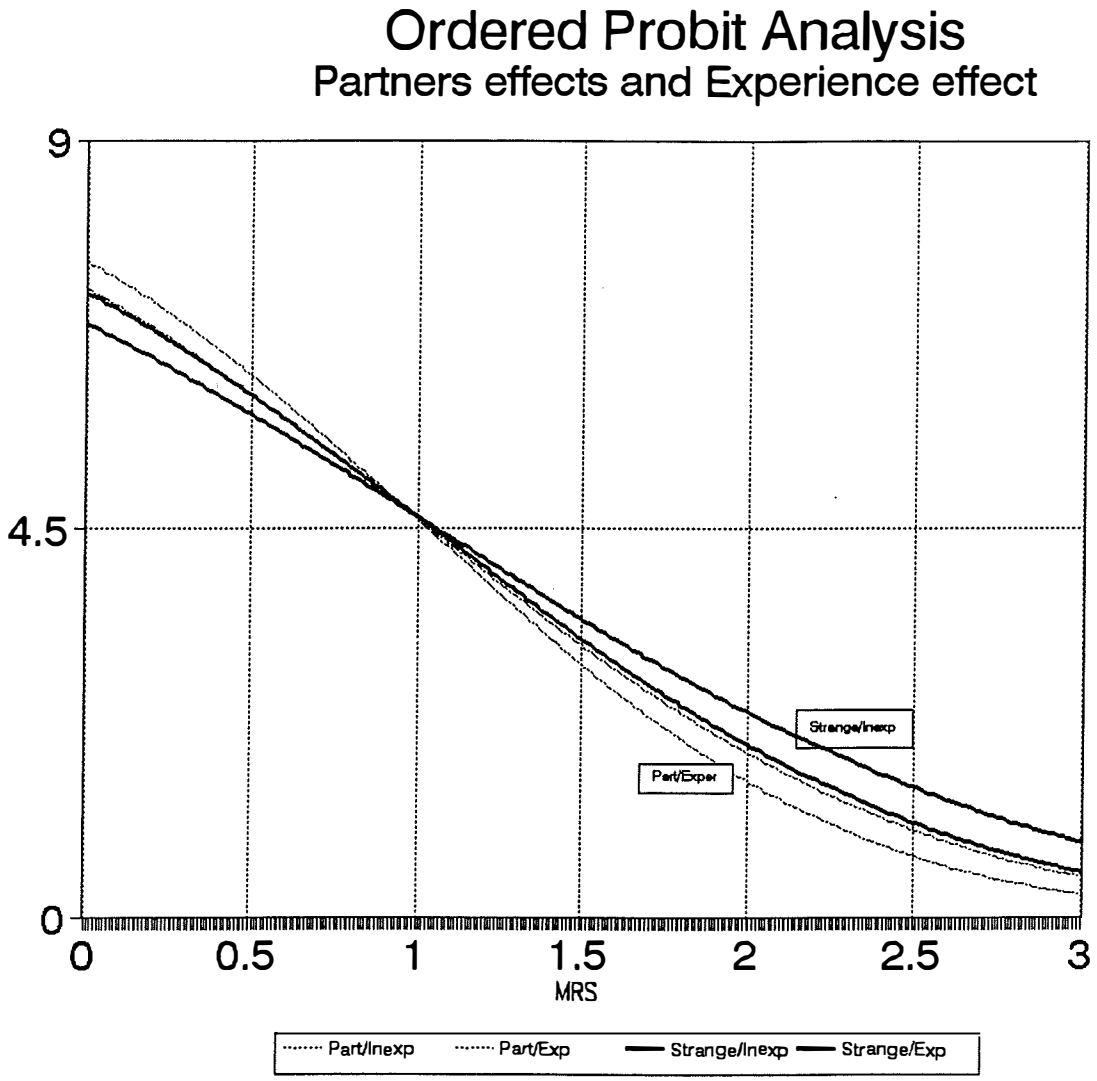
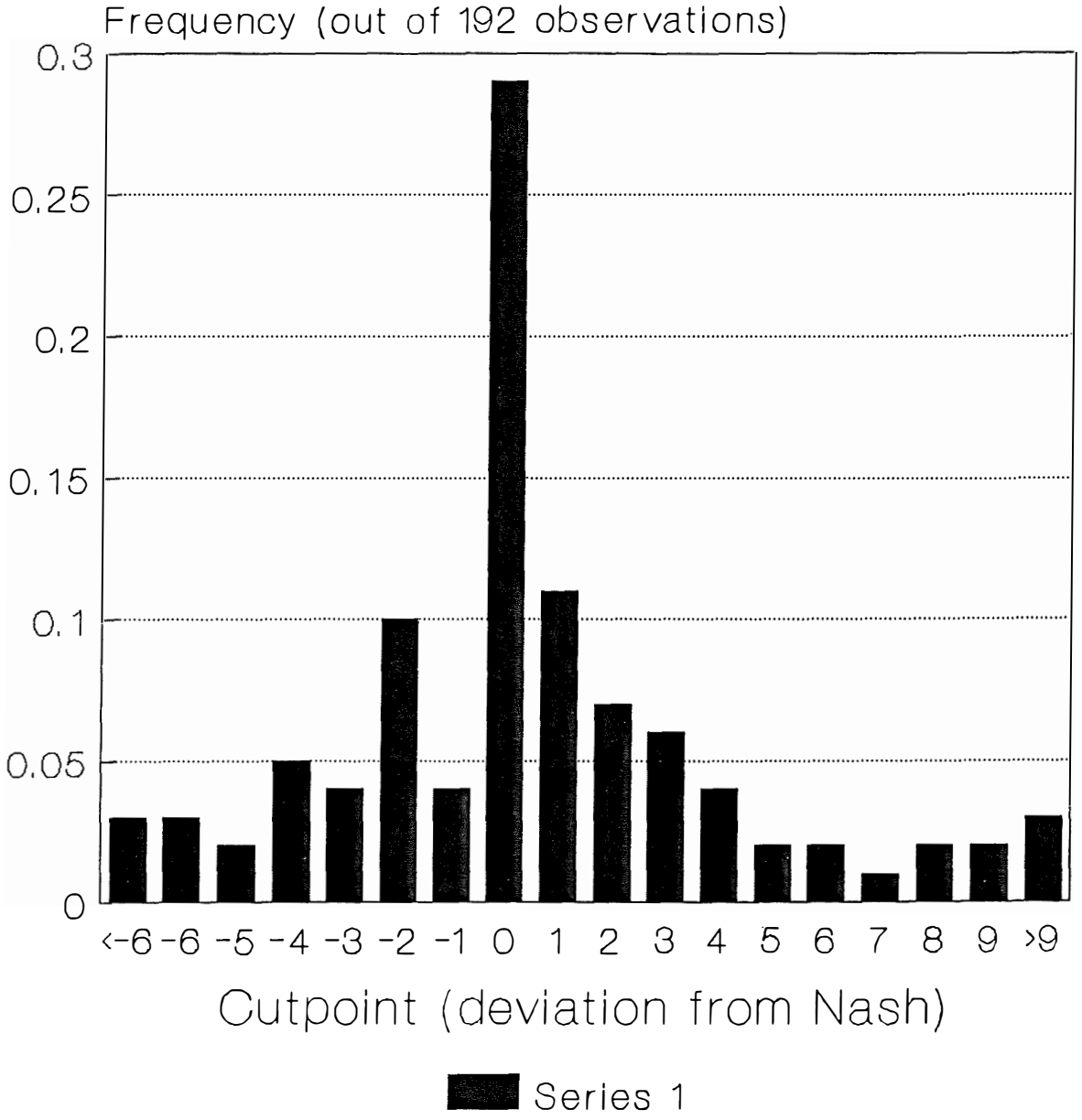


Figure 3: The expected contribution as a function of r/V as estimated by Ordered Probit Model.

Individual Cutpoints
UPF Data



Classification error minimizing cutpoints

Figure 4

Individual Cutpoints
UPF Partners vs. Strangers

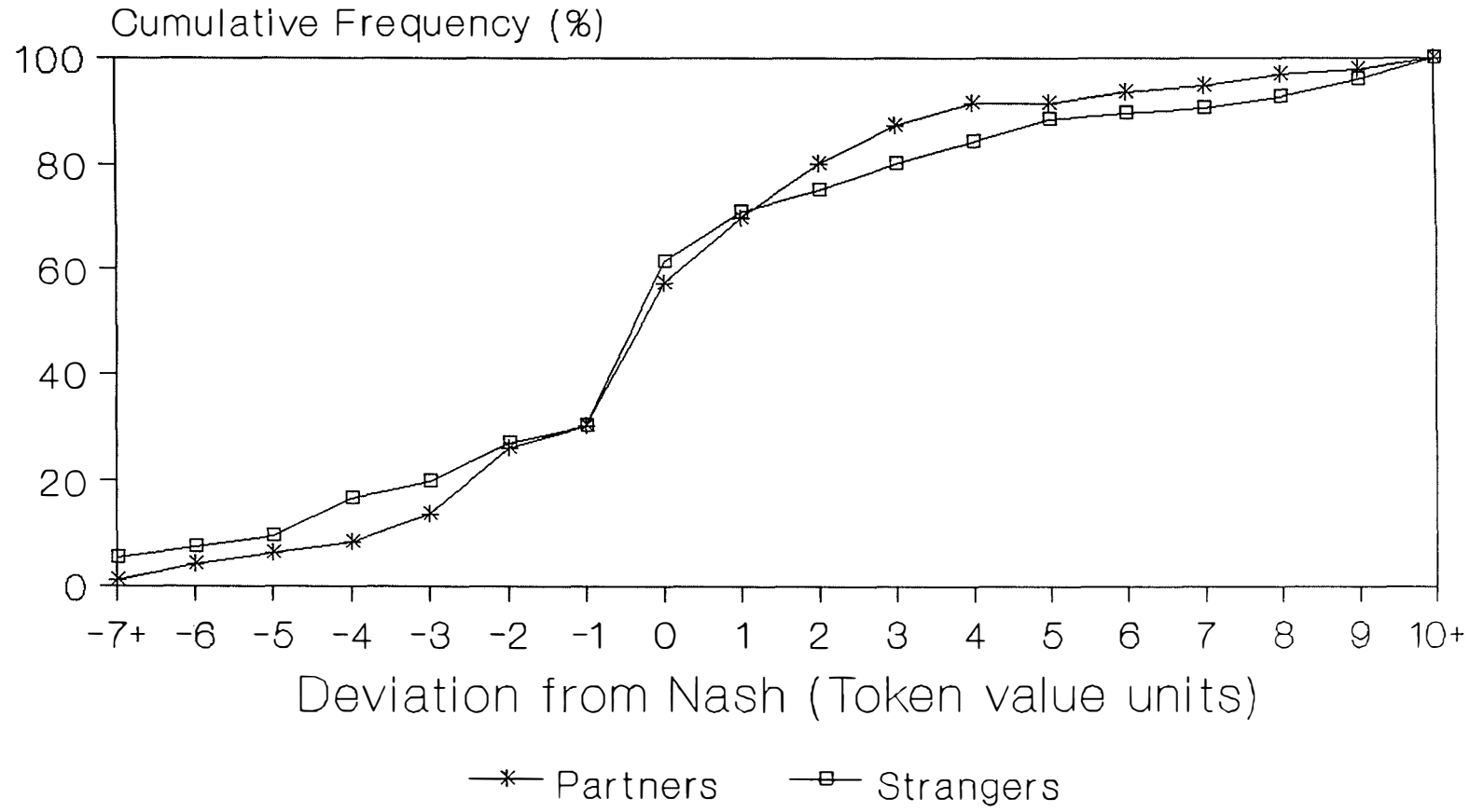
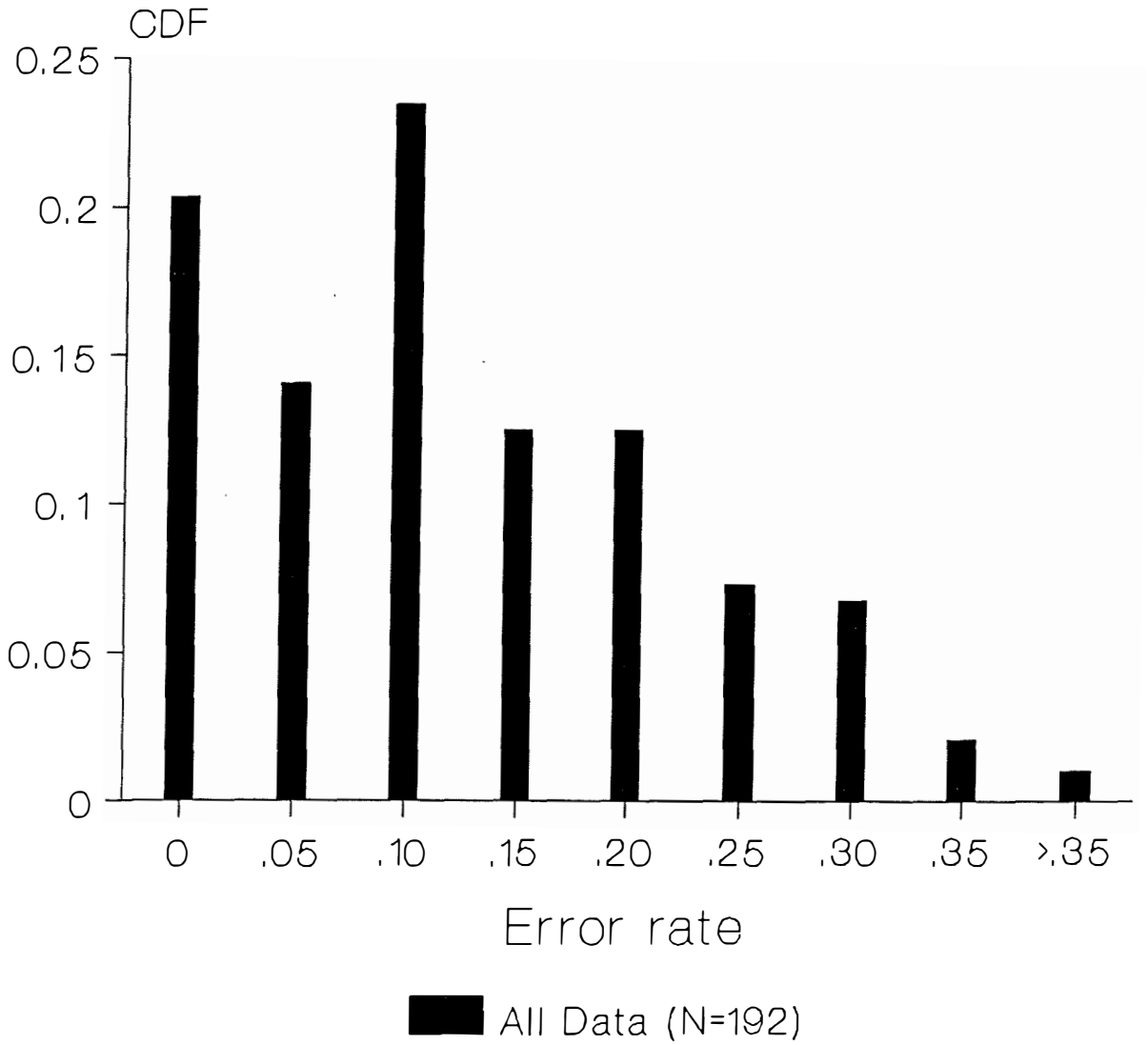


Figure 5
25

Cumulative Distribution of
Deviation from Nash Cutpoint

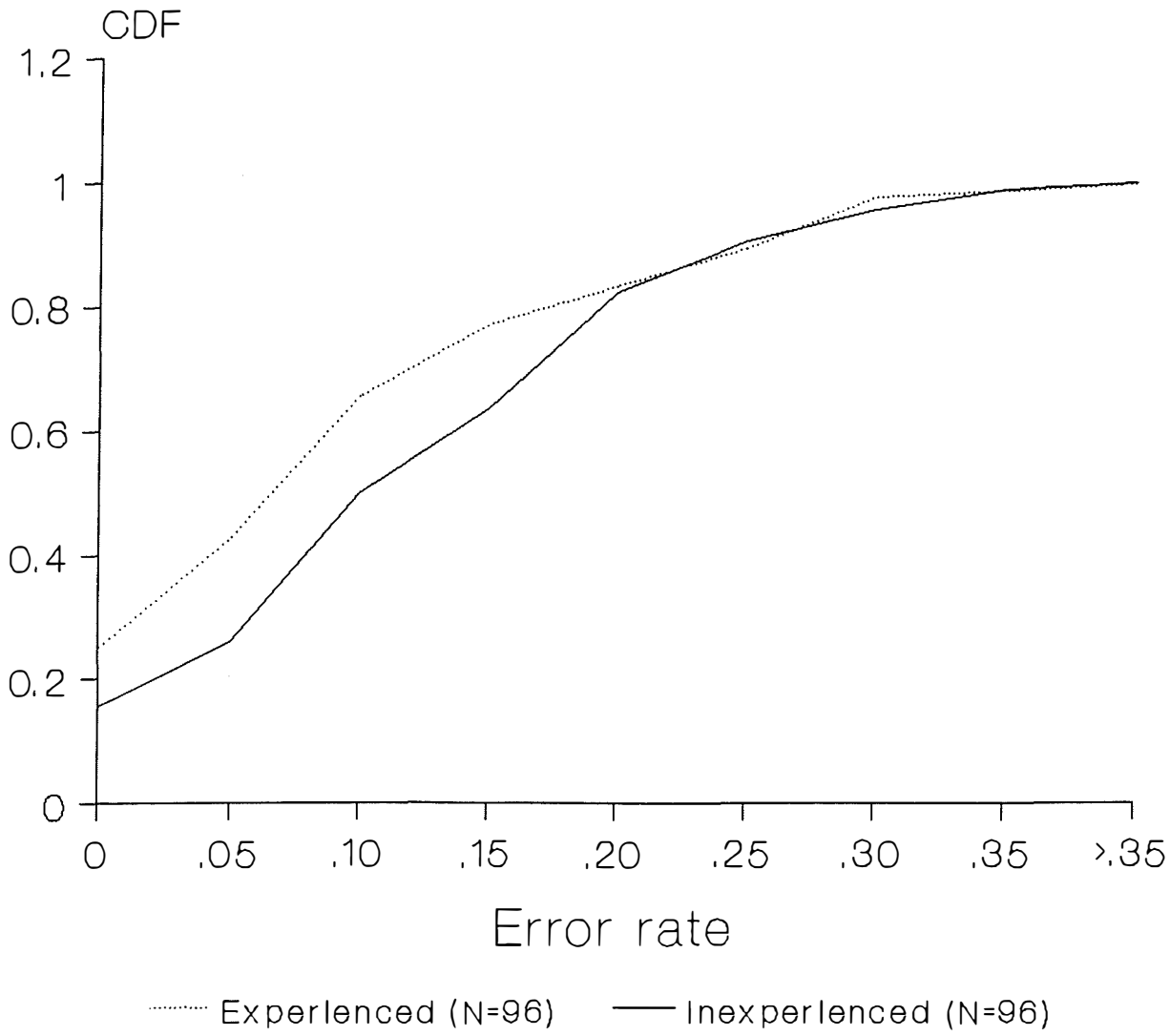
Errors
UPF Data



Fraction of decisions misclassified
relative to estimated cutpoint
Cumulative Frequence Distribution

Figure 6

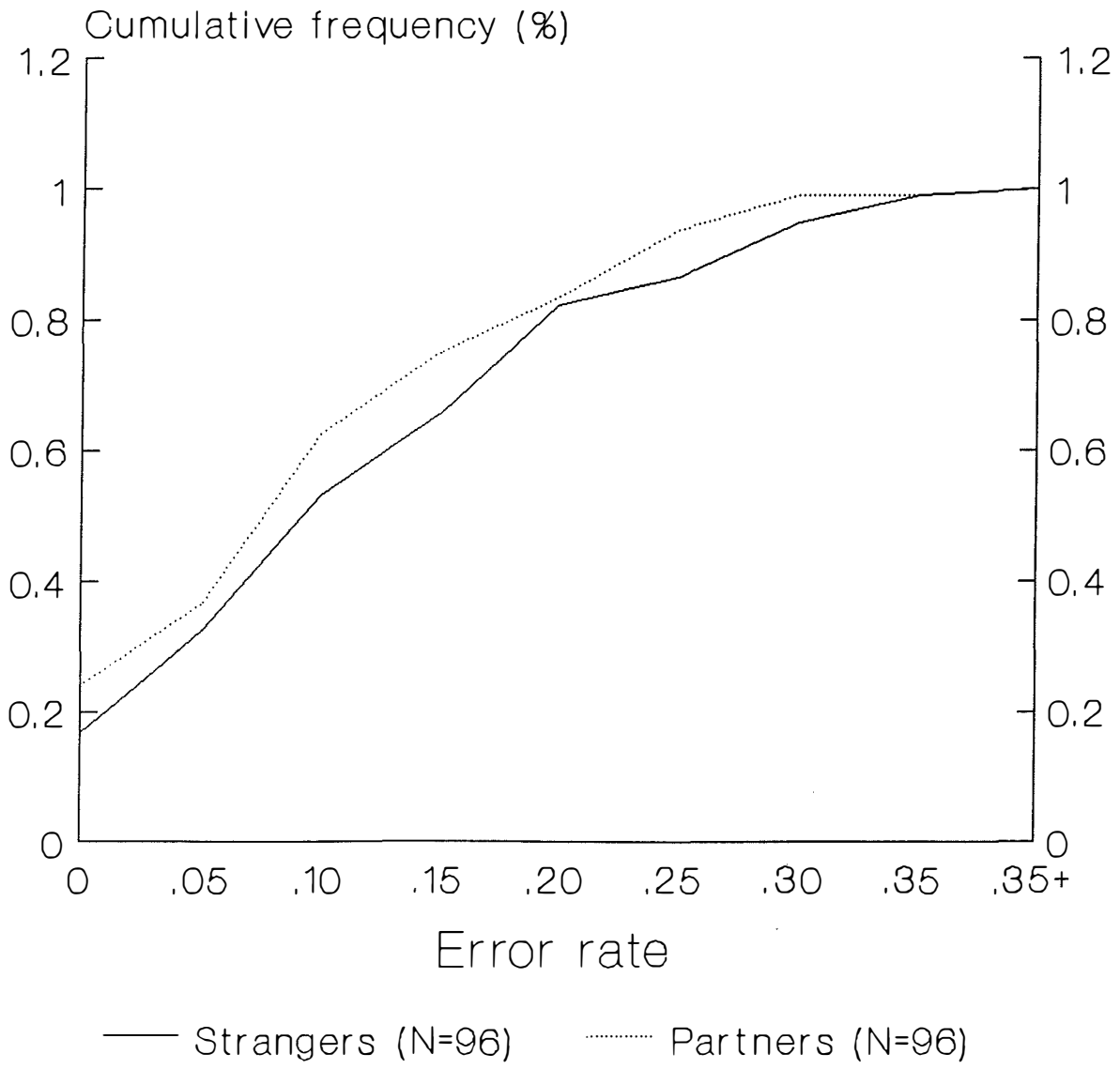
Errors
UPF Data
Experience effects



Fraction of decisions misclassified
relative to estimated cutpoint
Cumulative Frequence Distribution

Figure 7

Errors
UPF Data
Partners vs. Strangers



Fraction of decisions misclassified
(Distribution)

Figure 8

Errors/Cutpoints
UPF Data

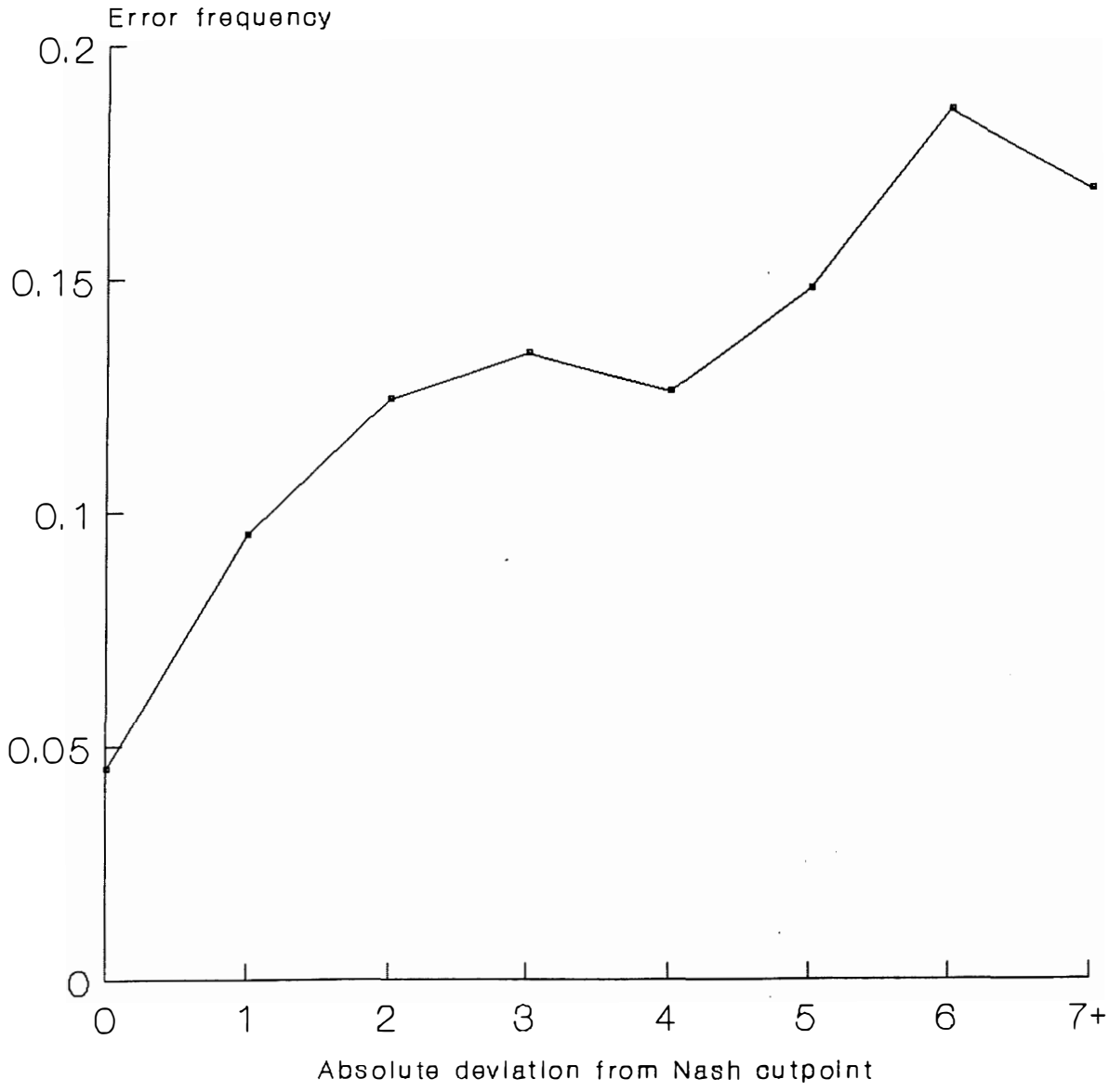


Figure 9

References

Andreoni, James. 1988. "Why Free Ride? Strategies and Learning in Public Goods Experiments." *Journal of Public Economics*, 37:291–304.

Andreoni, James. 1992. "Cooperation in Public Goods Experiments: Kindness or Confusion?" typescript.

Dawes, R. M. 1980. "Social Dilemmas." *Annual Review of Psychology*, 31:169–93.

Isaac, R. Mark, and James M. Walker. February 1988. "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism." *The Quarterly Journal of Economics*, 103:179–201.

Isaac, R. Mark, James M. Walker, and Susan H. Thomas. 1984. "Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations." *Public Choice*, 43:113–49.

Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory*, 27:245-52.

Kreps, David M. and Robert Wilson. 1982. "Reputation and Imperfect Information." *Journal of Economic Theory*, 27:253–79.

Ledyard, John, O. 1993. "Public Goods: A Survey of Experimental Research." To appear in *Handbook of Experimental Economics* (J. Kagel and R. Roth, eds.).

Marwell, Gerald and Ruth E. Ames. 1979. "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem." *American Journal of Psychology*, 84:1335–1360.

Marwell, Gerald and Ruth E. Ames. 1980. "Experiments on the Provision of Public Goods II. Provision Points, Stakes, Experience and the Free Rider Problem." *American Journal of Psychology*, 85:926–37.

Marwell, Gerald and Ruth E. Ames. 1981. "Economists Free Ride, Does anybody Else? Experiments on the Provision of Public Goods, IV." *Journal of Public Economics*, 15:295–310.

McKelvey, Richard D. and William Zavoina. 1975. "A Statistical Model for the Analysis of Ordered Level Dependent Variables." *Journal of Mathematical Sociology*, 4:103–20.

Palfrey, Thomas R. and Jeffrey E. Prisbrey. 1992. "Anomalous Behavior in Linear Public Goods Experiments: How Much and Why?" Social Science Working Paper No. 833, California Institute of Technology.

Palfrey, Thomas R. and Howard Rosenthal. 1988. "Private Incentives and Social Dilemmas: The Effects of Incomplete Information and Altruism." *Journal of Public Economics*, 28:309–32.

---------. 1991. "Testing Game-Theoritic Models of Free Riding: New Evidence on Probability Bias and Learning," in *Laboratory Research in Political Economy*, (T. Palfrey ed.) Ann Arbor: University of Michigan Press, p. 239–68.

Rapoport, Amnon. 1987. "Research Paradigms and Expected Utility Models for the Provision of Step-level Public Goods." *Psychological Review*, 94:74–83.

Saijo, T. and H. Nakamura. 1993. "The "Spite" Dilemma in Voluntary Contributions Mechanism Experiments," unpublished manuscript.